

SPAM SMS FILTERING USING MACHINE LEARNING


MUBASHIRA A

REG NO:MES20MCA-2029

PRODUCT OWNER : MR. VASUDEVAN T V



TABLE OF CONTENTS

1. INTRODUCTION
 2. MODULES
 3. METHODOLOGY
 4. RESULTS
 5. FUTURE ENHANCEMENTS
 6. DEVELOPING ENVIRONMENT
 7. PROJECT PLAN
 8. USER STORY
 9. PRODUCT BACKLOG
 10. SPRINT PLAN
 11. SPRINT ACTUAL
- 

INTRODUCTION

- A MAIN OBSTRUCTION IN ELECTRONIC COMMUNICATIONS IS THE VAST PUBLICIZING OF UNWANTED, HARMFUL MESSAGES KNOWN AS SPAM MESSAGES. LOTS OF TIME OF CLIENT IS BEING WASTED FOR SORTING APPROACHING THESE MESSAGES AND ERASING UNDESIRABLE CORRESPONDENCE, SO THERE IS A NEED FOR SPAM FILTERING SO THAT ITS OUTCOMES CAN BE REDUCED.
- THE MAIN AIM IS TO DEVELOPMENT OF SUITABLE FILTERS THAT CAN APPROPRIATELY FILTERING THOSE MESSAGES AND RESULTS IN A HIGH-PERFORMANCE RATE.
- IN THIS PROJECT, SPAM FILTERING AIMS TO DIFFERENTIATE BETWEEN SPAM AND AUTHORIZED MESSAGES. HERE, THE EVALUATION OF IT IS DONE BY USING A MACHINE LEARNING ALGORITHM NAMED SVM.
- MACHINE LEARNING ALGORITHMS, ESPECIALLY SUPPORT VECTOR MACHINE (SVM), CAN PLAY A MAJOR ROLE IN SPAM FILTERING.

Introduction



MODULES

- DATA COLLECTION
- TRAINING
- TESTING
- RESULT
- ADMINISTRATION

METHODOLOGY

IN THIS PROJECT, ONE OF THE MOST USED TECHNIQUES AS THE BASE CLASSIFIER TO OVERCOME THE SPAM PROBLEM IS THE SUPPORT VECTOR MACHINE. SVM CLASSIFY SPAM THAT IS TO DIFFERENTIATE BETWEEN SPAM AND AUTHORIZED MESSAGES. FIRST DATASET CONTAINING 5000 MESSAGES SAMPLES CONTAINING BOTH SPAM AND HAM(NON-SPAM) TYPE MESSAGES IS USED. WHERE 60% MESSAGES ARE USED FOR TRAINING AND REMAINING 40% ARE USED FOR TESTING. LATER VOCABULARY IS BUILT, WHICH CONTAINS A SET OF MOST FREQUENTLY WORDS CHOSEN FROM THE TRAINING/TESTING SET. VOCABULARY IS THEN USED TO CALCULATE THE TF-IDF(TERM FREQUENCY-INVERSE DOCUMENT FREQUENCY) VALUES, WHERE EACH MESSAGES WILL BE REPRESENTED ON THE BASIS OF THE IMPORTANCE OF WORD IN THE ENTIRE DATASET WHICH IS A N DIMENSIONAL VECTOR. THIS VECTOR IS FEATURE VECTOR. A MACHINE LEARNING ALGORITHM, SUPPORT VECTOR MACHINE(SVM) IS TRAINED TO CLASSIFY THE GIVEN MESSAGES. EACH OF THESE MESSAGES BELONGS TO ONLY ONE OF TWO CLASSES. THE IDEA OF SVM CLASSIFICATION IS FIND A LINEAR SEPARATION BOUNDARY THAT CORRECTLY CLASSIFIES TRAINING SAMPLES. AND LATER THIS MODEL IS USED TO PREDICT NEW MESSAGES GIVEN, WHICH IS THE MAIN AIM OR THE SYSTEM..

SPAM FILTERING PROCESS:

SMS SPAM IS MANUALLY CLASSIFIED INTO HAM AND SPAM MESSAGES WHICH ARE GIVEN AS THE INPUT OR THE TRAINING DATA FOR THE SPAM FILTER ALGORITHMS

STAGES:

❑ DATA COLLECTION:

COLLECT SPAM HAM DATASET FROM KAGGLE WEBSITE

❑ DATA PRE-PROCESSING:

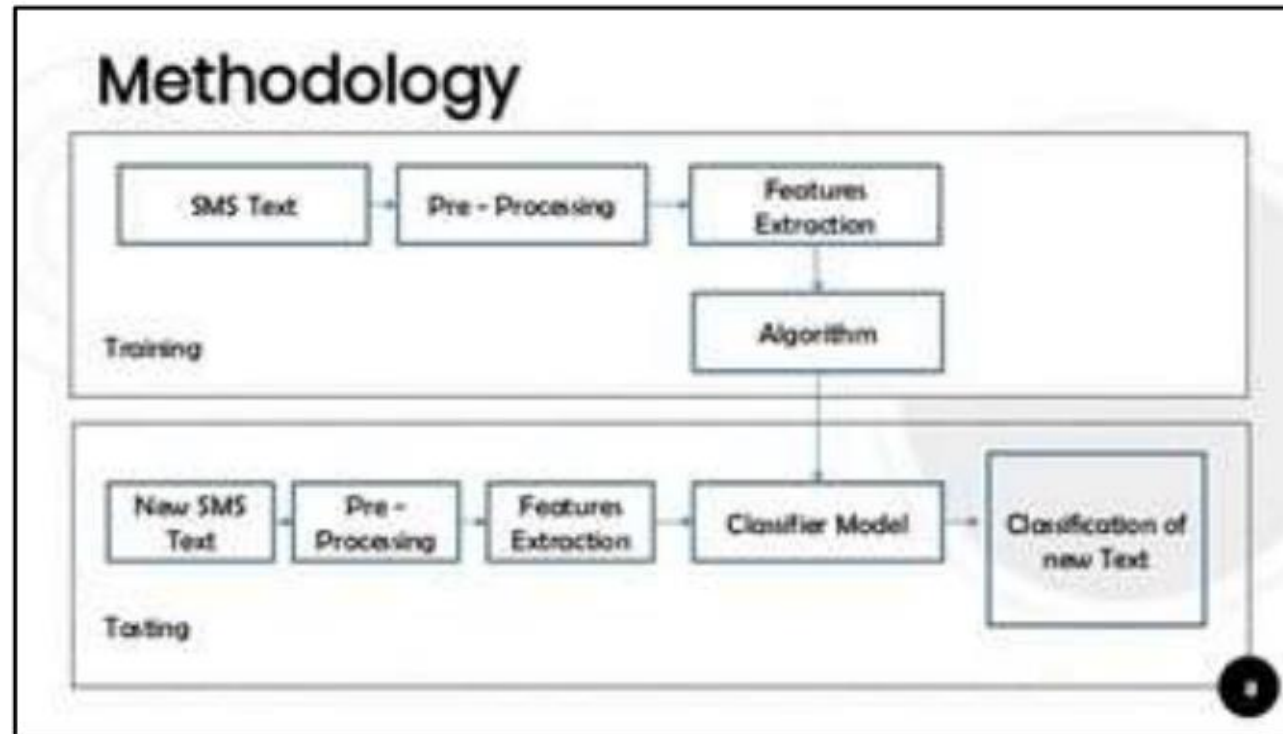
DATA PRE-PROCESSING IS USED TO TRANSFORM THE RAW DATA TO PREDICTABLE FORMAT. IN THIS STAGE, THE IMMATERIAL DATA SUCH AS STOP WORDS ARE ELIMINATED.

❑ TRAINING:

IT TRAINS THE ALGORITHMS FOR THE IMPORTANT ATTRIBUTE VALUES. WHERE 70% MESSAGES ARE USED FOR TRAINING. A MACHINE LEARNING ALGORITHM, SUPPORT VECTOR MACHINE(SVM) IS TRAINED TO CLASSIFY THE GIVEN MESSAGES.

❑ TESTING:

IT TESTS THE NEW DATA WITH THE TRAINING MODEL. WHERE 30% MESSAGES ARE USED FOR TESTING.



❑ RESULT:

PREDICT NEW MESSAGES IS SPAM OR NOT AND OUTPUT GENERATED.

ALGORITHM-SVM

SUPPORT VECTOR MACHINE, ABBREVIATED AS SVM, IS USED FOR EVERY TASK OF REGRESSION AND CLASSIFICATION. BUT, IT'S COMMONLY USED IN GOALS FOR CLASSIFICATION. THE AIM OF THE SUPPORT VECTOR MACHINE ALGORITHM IS TO SEARCH FOR A HYPERPLANE IN THE ASSOCIATED N-DIMENSIONAL SPACE (N — THE QUANTITY OF FEATURES) WHICH SEPARATELY CLASSIFIES THE INFORMATION POINTS. THERE ARE SEVERAL POTENTIAL HYPERPLANES WHICH WOULD BE CHOSEN TO SEPARATE THE 2 CATEGORIES OF INFORMATION POINTS. OUR GOAL IS TO SEARCH FOR A PLANE WITH THE UTMOST MARGIN, THAT IS, THE UTMOST DISTANCE BETWEEN THE POINTS OF KNOWLEDGE OF EACH CATEGORY. INCREASING THE MARGIN GAP SHOULD PROVIDE SOME REINFORCEMENT SO THAT FUTURE DATA POINTS CAN BE IDENTIFIED WITH EXTRA CONFIDENCE.

DEVELOPING ENVIRONMENT

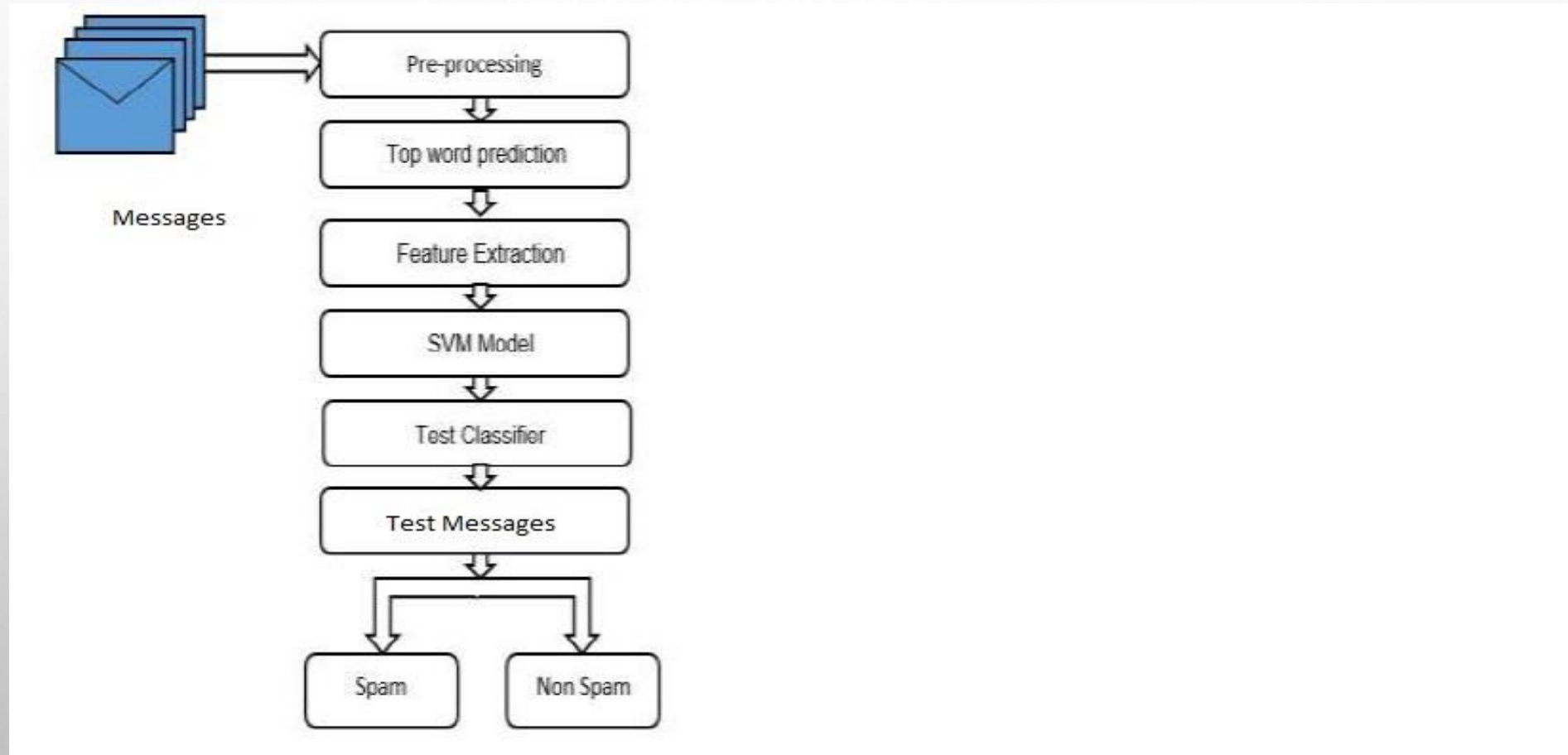
HARDWARE SPECIFICATION

- PROCESSOR : I3
- HARD DISK : 500 GB
- RAM : 8 GB

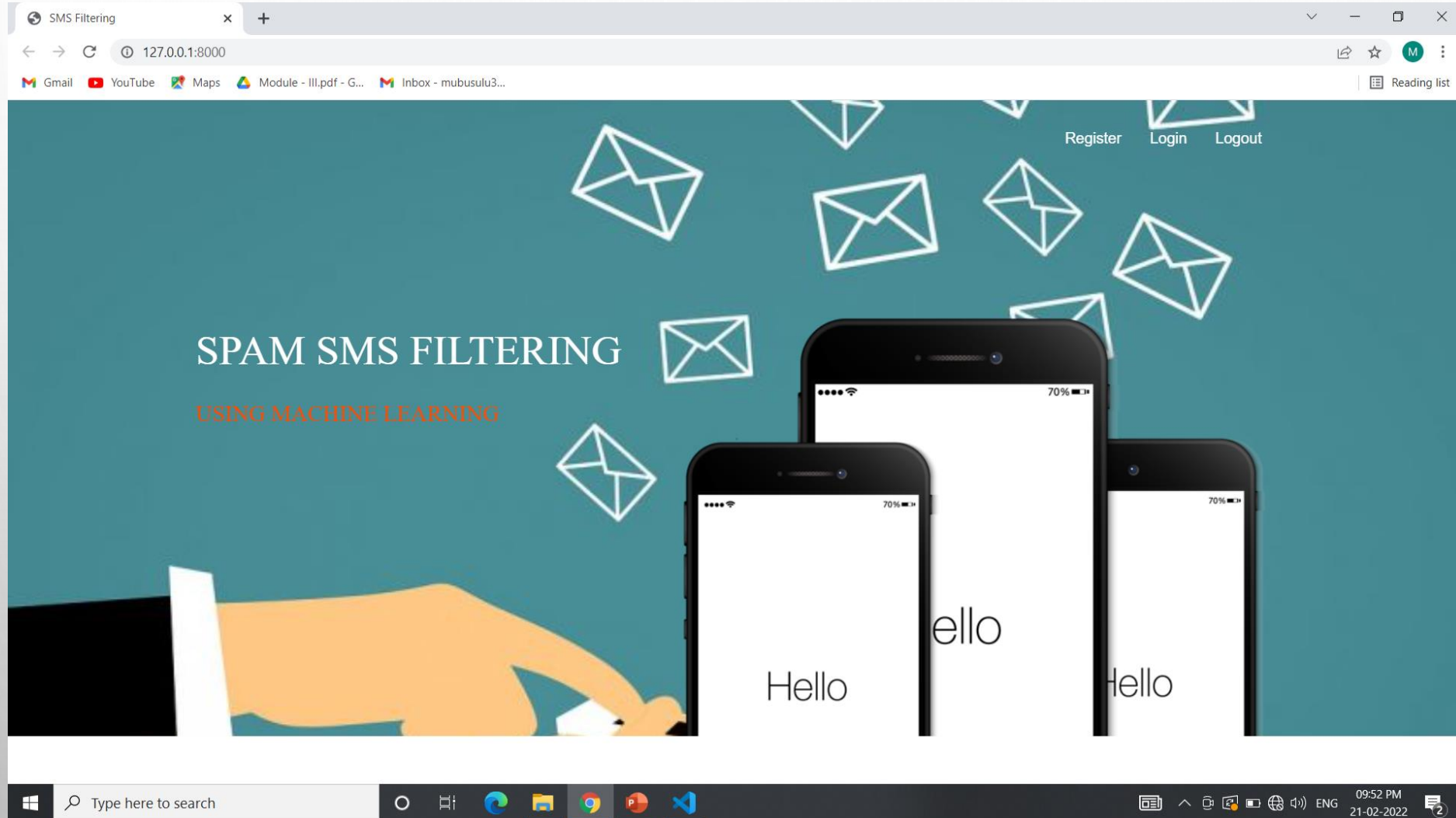
SOFTWARE SPECIFICATION

- LANGUAGE : PYTHON
- FRONT END : PYTHON-DJANGO
- BACK END : SQLITE
- DATASET : SPAM HARM DATASET FROM KAGGLE WEBSITE
- ALGORITHM : SVM
- IDE : VISUAL STUDIO CODE
- OS : WINDOWS

BASIC ARCHITECTURE



RESULTS



127.0.0.1:8000/register

127.0.0.1:8000/register

Gmail YouTube Maps Module - III.pdf - G... Inbox - mubusulu3... Reading list

Registration

Username:

Required. 150 characters or fewer. Letters, digits and @/./+/-/_ only.

First name:

Last name:

Email:

Password:

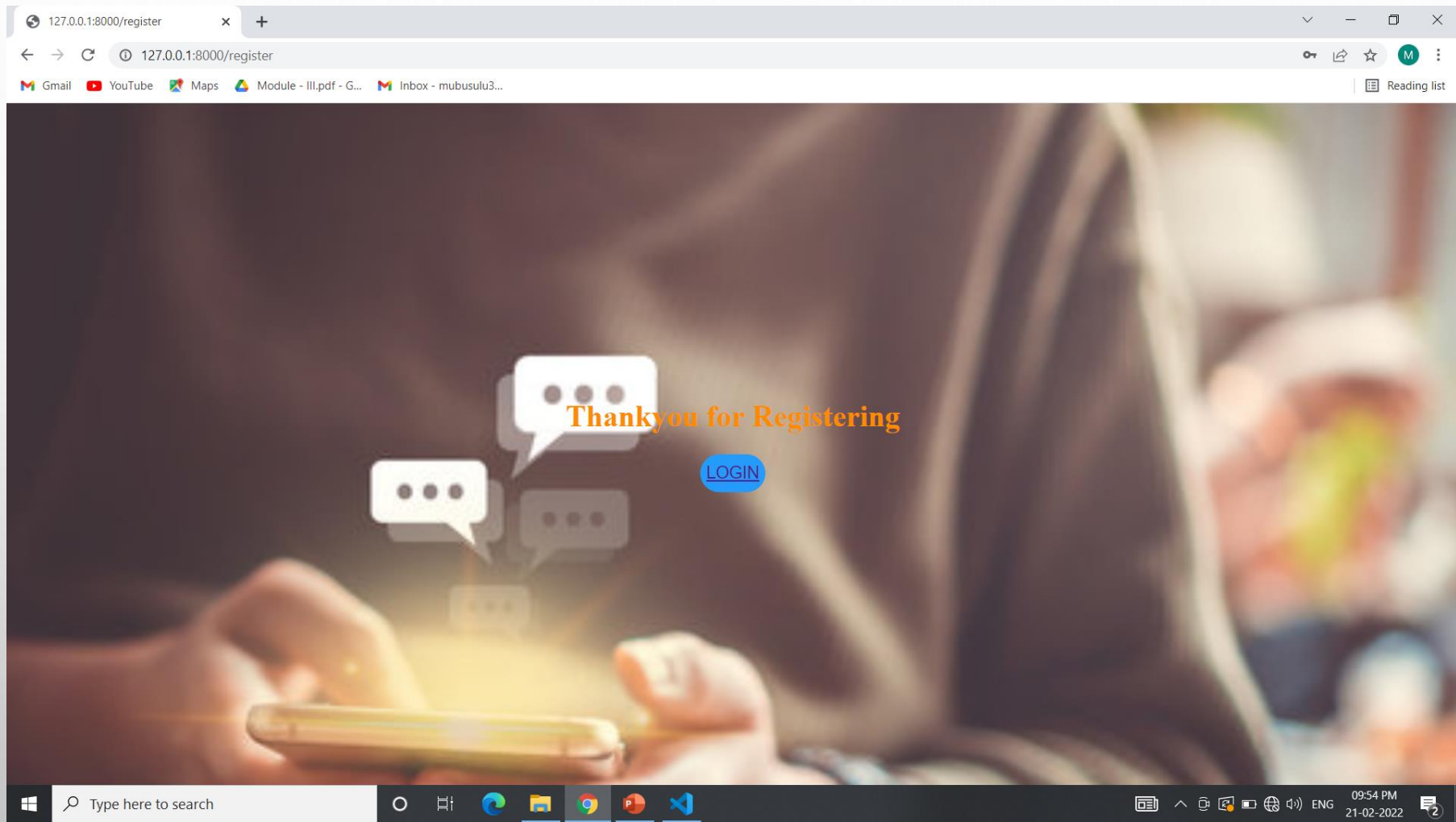
- Your password can't be too similar to your other personal information.
- Your password must contain at least 8 characters.
- Your password can't be a commonly used password.
- Your password can't be entirely numeric.

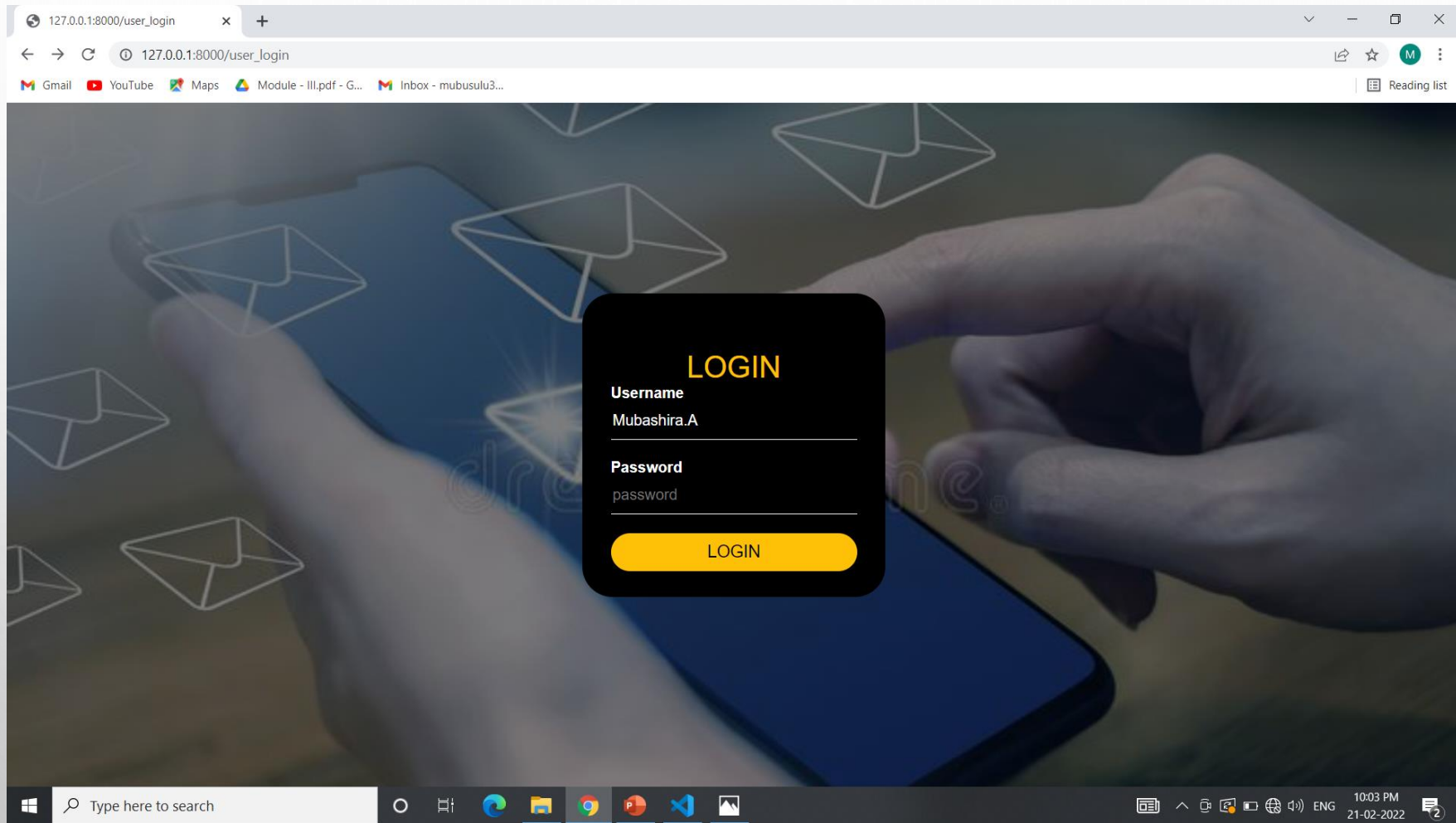
Password confirmation:

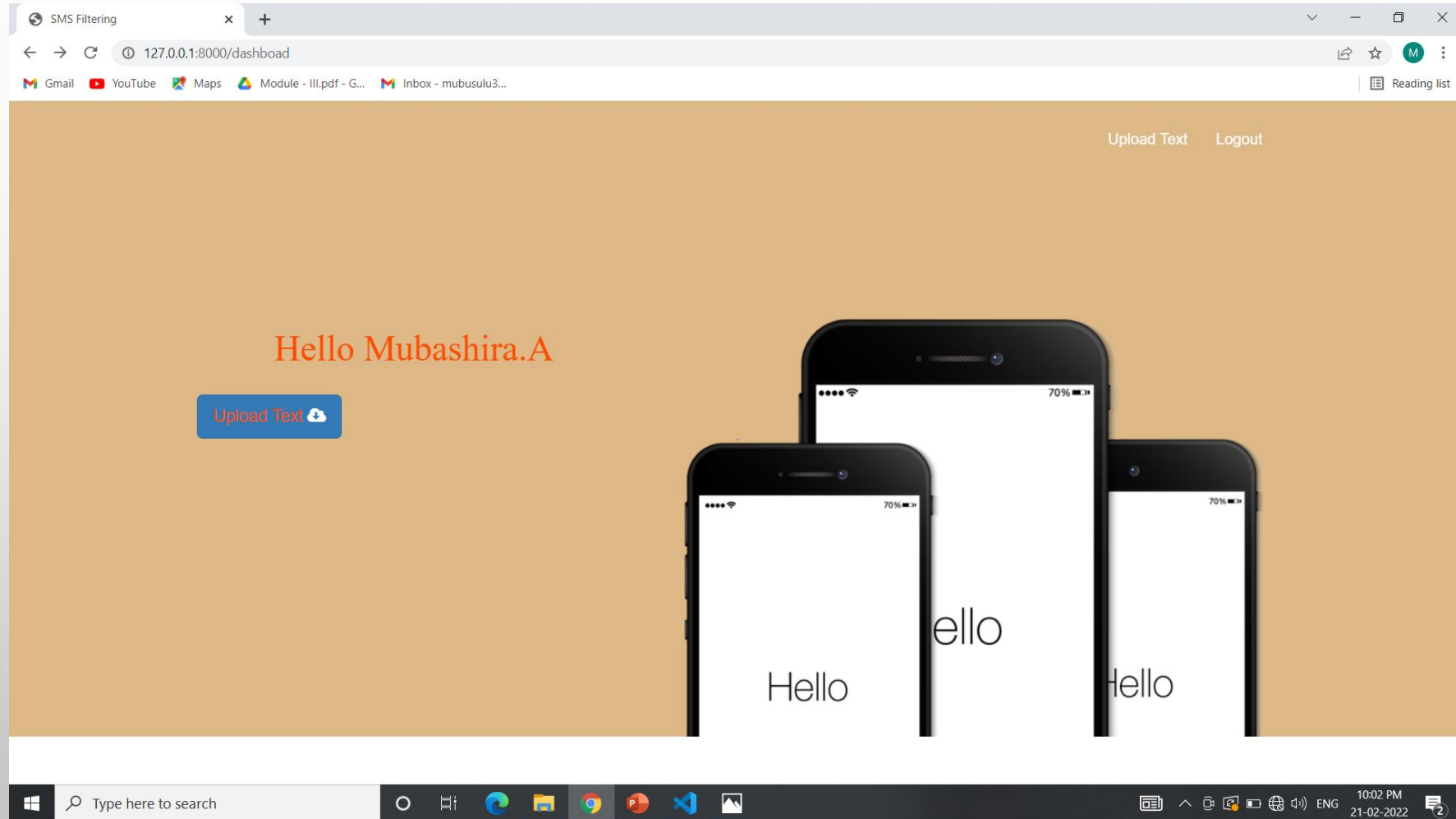
Enter the same password as before, for verification.

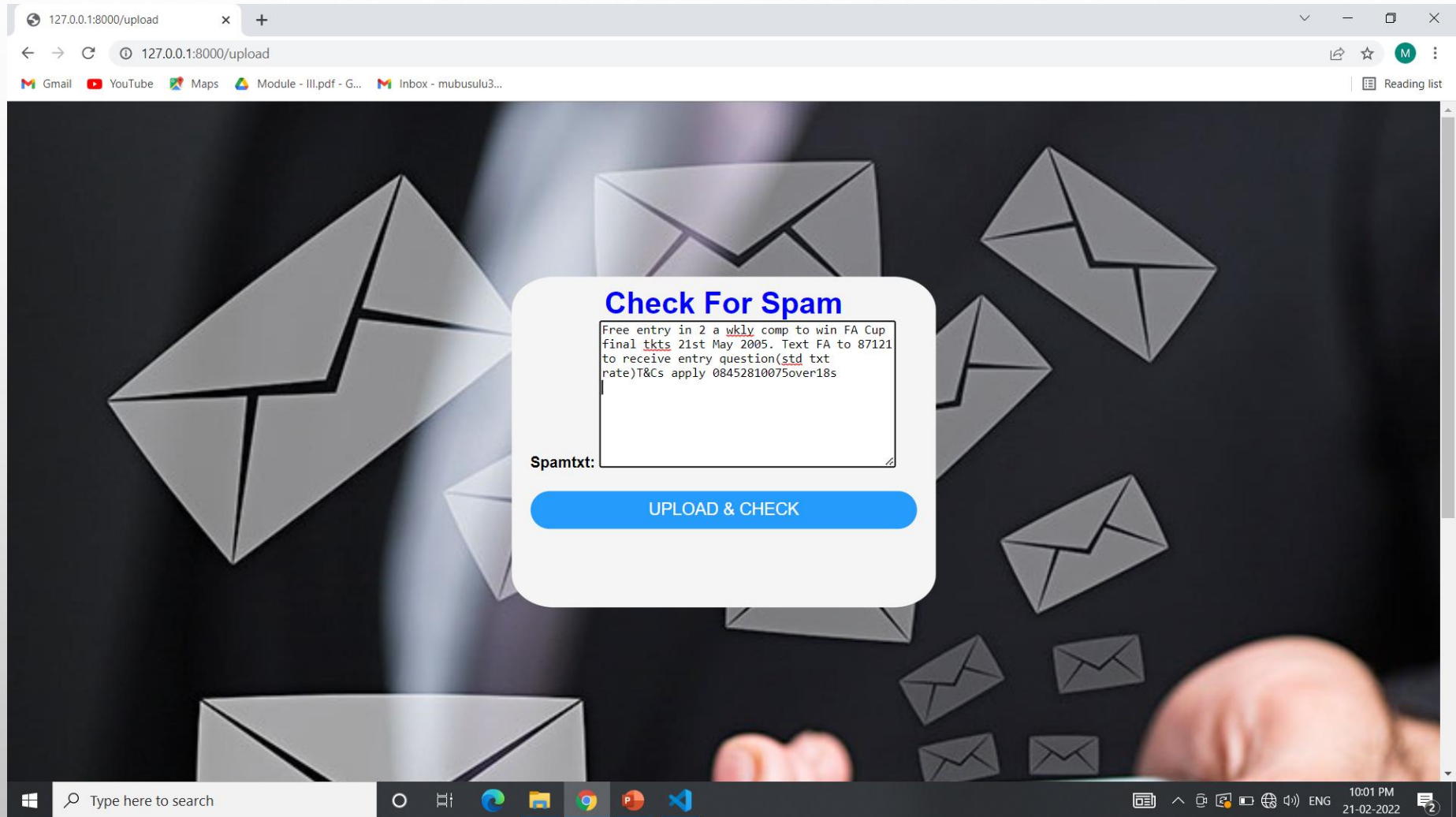
Type here to search

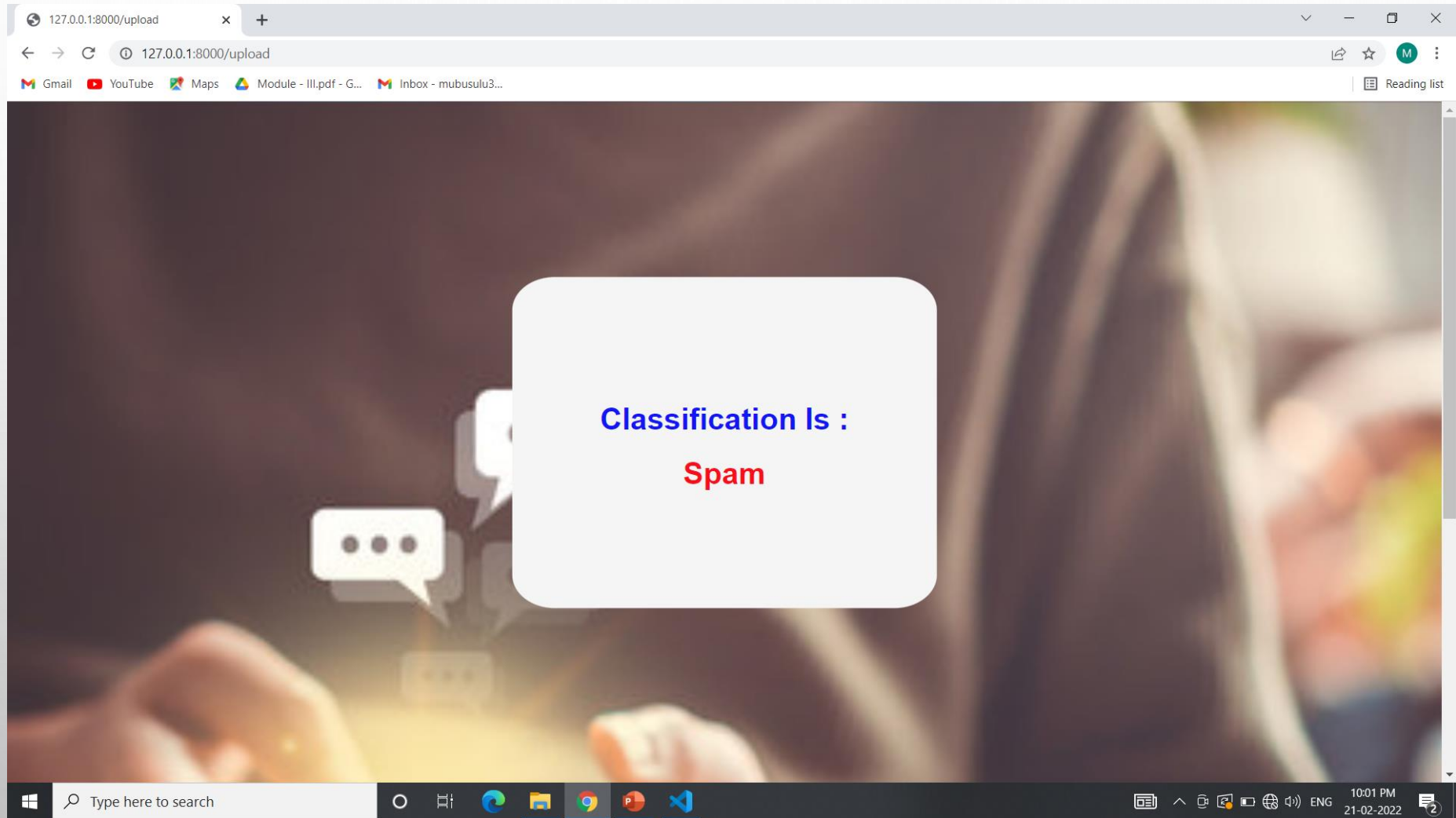
10:04 PM
21-02-2022











FUTURE ENHANCEMENTS

IN EXISTING SYSTEM THERE IS A FACILITY TO DETECT SPAM SMS , SO IN PROPOSED SYSTEM WE MODIFY SMS BY CATEGORIZING AS ADVERTISING , EDUCATION AND SPORTS ETC, AND WE DETECT WHETHER THE SMS IS SPAM OR HAM.

PROJECT PLAN

User Story ID	Task Name	Start Date	End Date	Days	Status
1	Sprint 1	30/11/21	30/11/21	2	Completed
2		15/12/21	15/12/21		Completed
3	Sprint 2	29/12/21	30/12/21	4	Completed
4		15/01/22	16/01/22		Completed
5	Sprint 3	22/01/22	23/01/22	4	Completed
6		29/01/22	30/01/22		Completed
7	Sprint 4	05/02/22	06/02/22	3	Completed
8		12/02/22	12/02/22		Completed

USER STORY

User Story ID	As a type of User	I want to <perform some task>	So that I can < Achieve Some Goal>
1	Admin	Login	I can get access to the system
2	User	Register	Registering users
3	User	Login	I can get access to the system
4	User	Result data collection form	Printing form
5	User	Training the data	View accuracy of the model
6	User	Modelling the data	Model created with SVM Algorithm
7	User	Testing the data	Input new dataset
8	User	Output generation	Output generated

PRODUCT BACKLOG

User Story ID	Priority < High /Medium /Low >	Size (Hours)	Sprint t <#>	Status < Planned / InProgress / Completed >	Release Date	Release Goal
1	High	5	1	Completed	30/11/21	Collect spam ham dataset from Kaggle website
2	Medium	5		Completed	15/12/21	Data pre-processing
3	Medium	5	2	Completed	29/12/21 - 30/12/21	User Interface Design
4	High	10		Completed	15/01/22 - 16/01/22	Result data collection form
5	High	7	3	Completed	22/01/22 - 23/01/22	Training the dataset
6	High	8		Completed	29/01/22 - 30/02/22	Apply SVM Algorithm and create model
7	High	8	4	Completed	05/02/22 - 06/02/22	Testing and predicting (whether it is spam or not)
8	High	2		Completed	12/02/22	Output generation

SPRINT BACKLOG PLAN

Backlog Item	Status and Completion date	Original Estimate in hours	Day 1	Day 2	Day 3	Day 4	Day 5	Day 6	Day 7	Day 8	Day 9	Day 10	Day 11	Day 12	Day 13
User story #1,2			Hours	Hours	Hours	Hours	Hours	Hours	Hours	Hours	Hours	Hours	Hours	Hours	Hours
Dataset Collection	30/11/21	5	5	0	0	0	0	0	0	0	0	0	0	0	0
Analysis	15/12/21	5	0	5	0	0	0	0	0	0	0	0	0	0	0
User story #3,4															
UI Designing	30/12/21	5	0	0	3	2	0	0	0	0	0	0	0	0	0
Coding	16/01/22	10	0	0	0	0	5	5	0	0	0	0	0	0	0
User story #5,6															
Training	30/01/22	15	0	0	0	0	0	0	3	4	4	4	0	0	0
User story #7,8															
Testing	12/02/22	10	0	0	0	0	0	0	0	0	0	0	4	4	2
Total		50	5	5	3	2	5	5	3	4	4	4	4	4	2

SPRINT 1 BACKLOG ACTUALS

Backlog Item	Status and Completion date	Original Estimate in hours	Day 1	Day 2	Day 3	Day 4	Day 5	Day 6	Day 7	Day 8	Day 9	Day 10	Day 11	Day 12	Day 13	Completed <Y/N>
User story #1,2			Hours	Hours	Hours	Hours	Hours	Hours	Hours	Hours	Hours	Hours	Hours	Hours	Hours	
Dataset Collection	30/11/21	5	5	0	0	0	0	0	0	0	0	0	0	0	0	Y
Analysis	15/12/21	5	0	5	0	0	0	0	0	0	0	0	0	0	0	Y
User story #3,4																
UI Designing	30/12/21	5	0	0	3	2	0	0	0	0	0	0	0	0	0	N
Coding	16/01/22	10	0	0	0	0	5	5	0	0	0	0	0	0	0	N
User story #5,6																
Training	30/01/22	15	0	0	0	0	0	0	3	4	4	4	0	0	0	N
User story #7,8																
Testing	12/02/22	10	0	0	0	0	0	0	0	0	0	0	4	4	2	N
Total		50	5	5	3	2	5	5	3	4	4	4	4	4	2	

SPRINT 2 BACKLOG ACTUALS

Backlog Item	Status and Completion date	Original Estimate in hours	Day 1	Day 2	Day 3	Day 4	Day 5	Day 6	Day 7	Day 8	Day 9	Day 10	Day 11	Day 12	Day 13	Completed <Y/N>
User story #1,2			Hours	Hours	Hours	Hours	Hours	Hours	Hours	Hours	Hours	Hours	Hours	Hours	Hours	
Dataset Collection	30/11/21	5	5	0	0	0	0	0	0	0	0	0	0	0	0	Y
Analysis	15/12/21	5	0	5	0	0	0	0	0	0	0	0	0	0	0	Y
User story #3,4																
UI Designing	30/12/21	5	0	0	3	2	0	0	0	0	0	0	0	0	0	Y
Coding	16/01/22	10	0	0	0	0	5	5	0	0	0	0	0	0	0	Y
User story #5,6																
Training	30/01/22	15	0	0	0	0	0	0	3	4	4	4	0	0	0	N
User story #7,8																
Testing	12/02/22	10	0	0	0	0	0	0	0	0	0	0	4	4	2	N
Total		50	5	5	3	2	5	5	3	4	4	4	4	4	2	

SPRINT 3 BACKLOG ACTUALS

Backlog Item	Status and Completion date	Original Estimate in hours	Day 1	Day 2	Day 3	Day 4	Day 5	Day 6	Day 7	Day 8	Day 9	Day 10	Day 11	Day 12	Day 13	Completed <Y/N>
User story #1,2			Hours	Hours	Hours	Hours	Hours	Hours	Hours	Hours	Hours	Hours	Hours	Hours	Hours	
Dataset Collection	30/11/21	5	5	0	0	0	0	0	0	0	0	0	0	0	0	Y
Analysis	15/12/21	5	0	5	0	0	0	0	0	0	0	0	0	0	0	Y
User story #3,4																
UI Designing	30/12/21	5	0	0	3	2	0	0	0	0	0	0	0	0	0	Y
Coding	16/01/22	10	0	0	0	0	5	5	0	0	0	0	0	0	0	Y
User story #5,6																
Training	30/01/22	15	0	0	0	0	0	0	3	4	4	4	0	0	0	Y
User story #7,8																
Testing	12/02/22	10	0	0	0	0	0	0	0	0	0	0	4	4	2	N
Total		50	5	5	3	2	5	5	3	4	4	4	4	4	2	

SPRINT 4 BACKLOG ACTUALS

Backlog Item	Status and Completion date	Original Estimate in hours	Day 1	Day 2	Day 3	Day 4	Day 5	Day 6	Day 7	Day 8	Day 9	Day 10	Day 11	Day 12	Day 13	Completed <Y/N>
User story #1,2			Hours	Hours	Hours	Hours	Hours	Hours	Hours	Hours	Hours	Hours	Hours	Hours	Hours	
Dataset Collection	30/11/21	5	5	0	0	0	0	0	0	0	0	0	0	0	0	Y
Analysis	15/12/21	5	0	5	0	0	0	0	0	0	0	0	0	0	0	Y
User story #3,4																
UI Designing	30/12/21	5	0	0	3	2	0	0	0	0	0	0	0	0	0	Y
Coding	16/01/22	10	0	0	0	0	5	5	0	0	0	0	0	0	0	Y
User story #5,6																
Training	30/01/22	15	0	0	0	0	0	0	3	4	4	4	0	0	0	Y
User story #7,8																
Testing	12/02/22	10	0	0	0	0	0	0	0	0	0	0	4	4	2	Y
Total		50	5	5	3	2	5	5	3	4	4	4	4	4	2	

The background is a light gray gradient. In the top-left and bottom-right corners, there are several realistic water droplets of varying sizes, some overlapping. A faint, circular, embossed-like pattern is visible in the upper center of the page.

THANK YOU